Communications Engineering Branch
Annual Report 2001
Submitted October 2001
George R. Thoma

The focus of the Communications Engineering Branch is applied research and development in image engineering and communications engineering motivated by NLM's mission-critical tasks such as document delivery, archiving, automated data entry for the creation of MEDLINE records, Internet access to biomedical multimedia databases, and imaging applications in support of medical educational packages employing digitized radiographic, anatomic, and other imagery. In addition to applied research, the Branch also developed and maintains operational systems for production of bibliographic records for NLM's flagship database, MEDLINE.

Areas of active investigation center on document image analysis and understanding techniques, image compression, image enhancement, image feature identification and extraction, image segmentation toward *query by image content* research, image transmission and video conferencing over networks implemented via asynchronous transfer mode (ATM) and satellite technologies, optical character recognition (OCR) and man-machine interface design applied to automated data entry. We also maintain the storage of large numbers of digitized spine x-rays and bit-mapped document images that are used for intramural and outside research purposes.

Information on these projects appears at http://archive.nlm.nih.gov/

Document imaging for the biomedical end-user

This research area has the goal of applying document image processing to document delivery via the Internet. Furthermore, it addresses NLM's mission of providing document delivery to end users and libraries, and incorporates advanced digital imaging techniques. The two active projects in this area are DocView and DocMorph.

DocView

First released in January 1998, this Windows-based client software has over 10,000 users in 144 countries. It facilitates the delivery of library documents directly to the patron via the Internet. Because DocView is compatible with Research Library Group's Ariel software, many biomedical libraries encourage their patrons to use it to receive, display, print and manage scanned images of journal articles and other

documents. While Ariel, a product of the Research Libraries Group, is used routinely by libraries and document suppliers to send documents via Internet to similar organizations, there are few options for *end users* to directly receive them.

The DocView client software, which runs under any version of Microsoft Windows, enables an end user to receive documents over the Internet at the desktop, retain them in electronic form, view the images, organize the received documents into "folders" and "file cabinets", electronically bookmark selected pages, manipulate the images (zoom, pan, scroll), copy and paste images, and print them if desired. DocView also serves as a TIFF viewer for compressed images received through the Internet by other means, such as web browsers. Users may receive document images either via Ariel FTP or Multipurpose Internet Mail Extensions (MIME) protocols. With DocView, users may also *forward* documents to colleagues for collaborative work.

DocMorph

The DocMorph server, in its second year of operation, has more than 3,000 registered users, many of whom are biomedical document delivery librarians. It serves as an important resource for librarians to convert library information from one form to another, often making it easier to exchange information. For instance, it is widely used to convert more than 50 different file formats to PDF for multi-platform delivery to patrons. By combining OCR with speech synthesis, DocMorph also enables the visually impaired to use library information. In August, Dr. Richard Smith, director of the Wolfner Library for the Blind and Physically Handicapped, reported using it to convert documents to synthetic speech recorded onto audio tapes for his blind patrons. To date, more than 29,000 jobs have been submitted to DocMorph, representing 335,000 pages of information consisting of 29 Gbytes of data.

The file types handled by DocMorph include: all file types from RLG's Ariel system, primarily for users receiving Ariel documents from their libraries; 9 TIFF types including uncompressed, G3, G4, monochrome and color; word processing including Word, Word for Mac, WordPerfect; Excel (.XLS) and PowerPoint (.PPT) files; JBIG and JPEG compressed files; DICOM; BMP; PSD, Adobe Photoshop; PCD, PhotoCD produced by digital cameras; and others.

In addition to these file types, considering the increasing popularity of DjVu, a new document compression scheme, an evaluation was conducted to possibly add DjVu compressed files as another candidate for conversion in DocMorph's repertoire. DjVu was compared against

other techniques: for monochrome document images, DjVu was more efficient than ITU G4 (by a factor of 3), TIFF JPEG (factor of 20) and TIFF LZW (factor of 100); for color images, DjVu was better than JPEG by a factor of 20. In addition to this testing, a dynamic link library that converts DjVu files to TIFF was developed. This DLL will be a building block to add DjVu file conversion to DocMorph.

The tools and subsystems used to achieve this include: inhouse imaging software written in C, C++ and assembly language; Kodak Image Edit control (available in the Windows operating system); MS Word to handle word processing file types; Image Magick, a freely available package; GhostScript, an Open Source package used with PostScript and PDF files.

While DocMorph's OCR facility converts TIFF images to text (and then to speech), there is no equivalent functionality for PDF and PostScript files. To extend DocMorph for this purpose, we integrated Ghostscript into DocMorph to implement a two step process: first converting PDF and PostScript files to TIFF by Ghostscript, and then have the OCR system convert the TIFF files to text and synthesized speech.

Based on research on DocMorph usage, a new client software is being designed. This software, MyMorph, will serve those who want a relatively hands-free method to convert a large number of files. DocMorph is useful for users with a small set of files to convert, but its *select file-upload-wait-process-wait-store file* technique is laborious and time consuming for users with a large number of files. The design is based on the new Simple Object Access Protocol (SOAP) technology that combines XML with HTTP to convert DocMorph into a Web service. With MyMorph, the user may select files for conversion to PDF, and the system will automatically perform the conversion over the Internet via DocMorph. Once converted, the files may be delivered to other users by MIME email. By providing the equivalent of a personal DocMorph server on the user machine, MyMorph is expected to increase user productivity.

The SOAP interface promises to enable DocMorph to serve as a building block facility serving as an information processing resource. SOAP will facilitate the integration of DocMorph in other systems and future projects such as file migration and language translation.

Document image analysis and understanding

DIAU research combined with database design, GUI design for workstations, image processing, speech recognition and related areas

underlie the development of MARS (Medical Article Records System), a system to automate the production of MEDLINE records from biomedical journals. The first generation of the MARS system was placed into operation, first at NLM, and in June 2000 moved to an offsite facility in downtown Bethesda. MARS-1, primarily a OCR-centered system designed to extract only the article abstracts while all other fields were manually entered, was supplanted by a second generation system (MARS-2) designed to extract the author names, affiliations and article title automatically. Performance data showed that while MARS-1 was a considerable improvement over the traditional keyboarding method, MARS-2 reduces the required labor effort to 25% of the manual approach.

Besides the automated subsystems (daemons) to extract those four fields, MARS-2 has four types of operator workstations: scanner, Edit, Reconcile and Admin. Edit is for entering information that is not extracted automatically; Reconcile is to verify the accuracy of the whole record prior to uploading it to the NLM database that retains records for subsequent indexing; and Admin for the operations supervisor to monitor the quality of the work and to direct the reworking of a process that appears improperly done.

Ongoing research in this area is described below:

Autozoning. The page segmentation stage following scanning blocks out ("zones") regions of contiguous text on the bitmapped image. Automatic zoning of contiguous text is a feature in the commercial OCR software used in MARS. However, tests showed that it was error-prone, and therefore would adversely affect the successful operation of critical downstream processes such as autolabeling (field identification) and autoreformatting. Its lack of reliability is related to its sole reliance on image data, and failure to exploit other output data from the OCR system. However, rather than eliminating this commercial autozoning feature entirely, we developed a method to exploit this feature. A four step process, combining both top-down and bottom-up strategies, is followed. First, the OCR output zones are disassembled into individual text lines. Then, the lines are split horizontally into fragments when word spaces exceed an empirically determined threshold. Third, the lines and line fragments are combined vertically into initial zones using as criteria vertical distance, line edge alignment and similarity of line features. Finally, these zones are combined into final zones using as criteria horizontal distance between initial zones, zone edge alignment and similarity of zone features. Our method was evaluated on 295 page images with 1180 zoned regions, and yielded an accuracy of 97.9%. It successfully corrected split zones, merged zones, and zones that are too small or too big for the text in the target fields (title, author, affiliation

and abstract). A daemon, therefore, was created for the autozoning function, and incorporated into MARS-2.

As more journals were added to the set to be processed by MARS, it was found that certain journals exhibited repeated problems with the zoning module in the Prime Recognition OCR, but not with another commercial OCR system from ScanSoft. We therefore incorporated an independent ScanSoft OCR module into the MARS system to handle these journals, on a journal-specific basis. The autozoning system is directed to use the zones from this module (and not from the Prime Recognition OCR) by journal ISSNs maintained in the MARS database.

Autolabeling. Identifying or labeling the zones of interest as *authors, title, affiliation and abstract* requires a family of autolabeling algorithms developed on the basis of a comprehensive set of 120 rules derived from both geometric as well as non-geometric (i.e., textual or numeric data) features from the OCR output. The geometric features include zone coordinates, zone height and width, average line height and length, average line spacing, and zone order. The non-geometric features include number of text lines in the zone, number of characters, number of punctuation marks, characters with particular font attributes (e.g., bold, italics, underlined), number of words, number of initials, strings representing academic degrees (M.D., Ph.D., etc.) average and maximum font size, and others.

The algorithms were tested against the images of articles from the journal titles indexed in MEDLINE excluding the approximately 1,000 titles for which publishers supply records in SGML form. The remaining 3,000+ titles are therefore candidates for the automatic processes in MARS. Errors encountered in testing the baseline algorithm were largely in labeling affiliation zones, and these were due to incorrect font attributes in the output of the commercial OCR system. To date, 2,028 journal titles can be processed automatically, but for 580 of these the publishers are delivering citations via XML tagged format, leaving 1,448 titles suitable for MARS-2 processing. This effort will continue until all the scanned journals are tested and the rules are tailored to allow automated processing of the largest possible number.

Affiliations correction. Affiliations offer a particular challenge to automated recognition since they are often printed in small characters and italics, predictably harder for an OCR system to convert correctly. The approach taken, lexical analysis, began with a detailed investigation of several string pattern matching techniques. A C++ module PatternMatch was developed that combines whole-word

matching (that is fast) with probability matching (that is more accurate, but slow). This cascaded process first performs whole-word matching with a small dictionary containing words with frequency of occurrence of 100 or more, followed by the inhouse-developed probability matching technique for those words not found in the first step. The dictionary for this second step is a larger one, including words that occurred 2 or more times. Experiments proved that the first step is fast, processes 45% of the low-confidence words, and has a false positive rate of only 1.4%. The remaining 55% of the low-confidence words are then processed by the probability matching algorithm. Our experiments showed that the overall results of this cascaded process is correct detection rate of 83.4%, false positive rate of 13%, and no matches 3.6%.

In practice, when the Reconcile operator clicks on a word in the affiliation field that is wrong or undecipherable, PatternMatch offers a drop-down word list in which the first word is usually the correct one. When that word is selected, it automatically replaces the incorrect word, and the operator does not have to type it in. In production, PatternMatch has presented the operators with thousands of word lists to date; the *first* word in these lists was correct 85.6% of the time, another word in the list was selected 8.9% of the time, and no correct word appeared in the list 5.5% of the time, closely matching the results of the earlier experiments.

Further research continues with techniques that use both zip codes as well as author names to match with affiliations. To set the stage for these investigations, 15,000 correct completed citation records (from existing MARS output data) have been collected, grouped by date, diacritics changed to the current conventions used by NLM, and files have been prepared to be accessible by browser. Also the dictionary used by PatternMatch has been enlarged to include institutional names that are hyphenated, not in the present dictionary.

The approach to use author names (that are usually recognized correctly) as a clue to the affiliation is under way. To determine feasibility, a MS Access database of 300,000 entries containing author/affiliation data from past MARS production was created, and an analysis of the distribution of numbers of articles associated with a given author indicates that 40% of incorrectly recognized affiliations might be correctable in this manner. This would be followed by the development of a string matching algorithm that matches words within strings to determine the degree of similarity. This algorithm would be used by MARS developers to create a DLL that generates similarity scores for past recorded (correct) affiliations corresponding to a particular author, so that the Reconcile operator could click select the

correct entry.

Edit workstation. Since manual entry of certain fields is required, the workstation that enables operators to do this has been improved in the following ways: (a) In the page image displayed to the operator, the labeled zones are tagged with a number indicating the percentage of high confidence characters. The purpose is to alert the operator to poor image-to-text conversion by the OCR so that a repeat of the scanning, the OCR operation or re-labeling of zones may be done. This feature relies on extracting zone and labeling information from the database, and the Kodak imaging tool to paint the numbers onto the image displayed. We are investigating the feasibility of identifying a threshold for the percentage figure to suggest that the journal issue be reprocessed. (b) Code was added to warn the operator, in the event the journal is one that is "non-compliant" to check the zones and labels before proceeding, since the automated processes may not correctly handle such journal titles. (c) The Edit operator is no longer required to key in author, title, etc for non-compliant journals, since we are assuming that the automated processes will correctly extract these for some of these journals.

CheckIn module. This subsystem has turned out to save a significant amount of the operator supervisor's time, eliminating the need to manually create a form containing special instructions and information related to an incoming journal issue, paper-clip it to the issue, and pass this on to the scan, Edit and Reconcile operators. The way this works: a journal arriving at the MARS supervisor's station is "wanded" (barcoded MRI number scanned in) into CheckIn; this MRI identifier is passed, as a parameter in a URL, to the DCMS website across the wide area network at NLM; this on-demand query returns XML information about the specific issue (volume number, publication date, etc.) and its series (ISSN, title, etc.); CheckIn now uses the ISSN to retrieve other series information from the MARS database; finally, all this data is parsed and printed out on a form that the supervisor paper-clips to the issue, and is subsequently referred to by the operators. If it turns out that the DCMS database does not return an ISSN, then CheckIn automatically attempts to find a match by title in the MARS database. In case this is unsuccessful, a wizard is immediately invoked explaining the problem and providing the means to look up relevant information from the MARS database by title or by ISSN.

In March, CEB research staff with scientists from Bell Labs and the University of Oulu, Finland, organized and presented a workshop on *Document Image Analysis and Understanding* at the First International NAISO Congress on Information Science Innovations (ISI 2001). This conference was held at the American University in Dubai, United Arab

Emirates.

Biomedical imaging and multimedia database R&D

The aim of this program is to address fundamental issues in the handling, organization, storage, access and transmission of very large electronic files in general and digitized x-rays in particular. A special focus is research into these topics as applied to heterogeneous multimedia databases consisting of both images and text. This work has evolved from a previous project named DXPNET for Digital Xray Prototype Network conducted in collaboration with two other agencies, the National Center for Health Statistics (NCHS) and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS).

Biomedical image-related work in CEB this year consisted of (1) the continuing development and beta testing of the WebMIRS system; (2) the continuing development and beta testing of the Digital Atlas of the Spine; (3) the enhancement of an online, publicly accessible archive of 17,000 digitized x-ray spine images; and (4) research into computer-assisted methods for the extraction of biomedical information directly from the spine images (for image indexing and query by image content.)

WebMIRS. The Web-based Medical Information Retrieval System is a Java applet that allows remote users to access data from two surveys conducted by the National Center for Health Statistics: the second and third National Health and Nutrition Examination Surveys (NHANES II and III), carried out during the years 1976-1980 and 1988-1994, respectively. The NHANES II database accessible through WebMIRS contains records for about 20,000 individuals, with about 2,000 fields per record; the NHANES III database contains records for about 30,000 individuals, with more than 3,000 fields per record. In addition, a user query may retrieve any of the 17,000 x-ray images collected in NHANES II, and display it in low-resolution form. WebMIRS allows a user to control a graphical user interface to construct a query of the NHANES II or NHANES III data. A sample query might be equivalent to the English statements: "Find records for all individuals who reported chronic back pain. Return their age, sex, race, age when the pain began, and longest duration of pain. Also, return the record data required for statistical analysis and display their x-ray images." WebMIRS allows the user to save the returned data to the local disk drive, for analysis by appropriate statistical tools such as the commercially available SAS and SUDAAN software. In effect, WebMIRS serves as a research tool by going beyond data access and retrieval to data analysis.

This year vertebral boundary data was added to the WebMIRS NHANES II database and made available for public use. The vertebral boundary data, produced by a board-certified radiologist for 550 of the 17,000 x-ray images in WebMIRS, consists of (x,y) coordinates for approximately 20,000 points on the vertebral boundaries in the cervical and lumbar spine images. This data is both to enable a new type of query (that exploits both radiological and health survey data) as well as to serve as ground truth data for image analysis. An example of this new type of query: "Find records for all persons having low back pain (health survey data) and fused lumbar vertebrae (radiological data)". The boundary data points are displayable on the WebMIRS image results screen and may be saved to the user's local disk. Current work is being done to support more advanced query and display using this radiological data.

Also this year, a new procedure to respond to a new WebMIRS user with an e-mail of detailed technical support options was incorporated; the Web pages for access to WebMIRS were revised for ease of use; and a test platform for evaluating WebMIRS under Windows 2000 was established.

A new installation method for WebMIRS, based on Java WebStart technology, is being developed to simplify the current process. This new method will replace the current process, which requires a user to install the Java Runtime Environment, install the WebMIRS security file, and then access the WebMIRS applet.

WebMIRS version 1.0.7b is being beta tested at 80 institutions at present, both within and outside the U.S. The last five to request access to the system: (i) Dept. of Information Science, Saga University, Saga Shi, Japan; (ii) Dept. of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC; (iii) Lata Medical Research Foundation, Nagpur, India; (iv) Dept. of Global Imaging Technology, GE India Technology Center, Bangalore, India; and (v) Mahidol University, Phutamonton, Thailand.

WebMIRS was used in two semesters of a graduate course in public health statistics at Columbia University in 1999-2000 to demonstrate new technological data access methods, and a real time data acquisition and analysis was demonstrated using WebMIRS/SAS/SUDAAN at the CDC Data Users Conference.

The Digital Atlas of the Spine. This is a dataset of cervical and lumbar spine images with interpretations validated by a consensus of medical experts, along with software to display and manipulate the images. The images in the Atlas were chosen from the 17,000 images in the

NHANES II survey. We convened two workshops in collaboration with other National Institutes of Health researchers to seek expert advice and consensus on a wide set of technical and biomedical issues related to the radiological interpretation of this set of images. Among the issues covered were the exact features to be interpreted. Radiographic features considered for interpretation of the cervical images were anterior osteophytes, posterior osteophytes, disc space narrowing, sclerosis, vacuum phenomenon, and subluxation. For the lumbar images, features considered included anterior osteophytes, posterior osteophytes, disc space narrowing, sclerosis, vacuum phenomenon, spondylolisthesis, spondylolysis, and DISH. A subset of these features was selected as likely to be consistently interpretable from the NHANES images. This selection of features, based on the consensus of experts at the workshop, took into account published studies relating to the likelihood of obtaining consistent readings for the features considered. The features identified by the workshop as consistently readable were those chosen for the Atlas.

For the cervical spine images, the Atlas contains numeric interpretations or "grades" for anterior osteophytes and disc space narrowing, on a scale from 0-3, with 0 being "normal" and 3 being "most abnormal"; and also interpretations for subluxation, on a 0-1 scale, with 0 being "normal" and 1 being "abnormal". Similarly, for the lumbar spine images, the Atlas contains interpretations for anterior osteophytes and disc space narrowing, on a scale from 0-3. The Atlas user may display single or multiple images in order to view, for example, all grades from normal to most abnormal of anterior osteophytes in the cervical spine. Image processing capability is provided to assist in contrast enhancement for viewing of detail.

The Atlas may be accessed either as a Java applet, or downloaded as a Java application, from the CEB website. In addition, we provide a version of the Java application on CD. The Java application version allows the user to add his/her own images in a special "My Images" section, and to annotate and title those images for later use.

This year the Atlas was redesigned with future applications in mind. It now allows the inclusion of color images into the "My Images" section for possible use in displaying endoscopic or other color images. We are investigating the potential for the Atlas to assist in the process of creating endoscopy reports that conform to standard medical terminology.

Version 2.0 of the Atlas is now being distributed for beta testing. Though started and led by research engineers at CEB, our collaborators are experts from the National Institute of Arthritis and Musculoskeletal

and Skin Diseases (NIAMS), the NIH Clinical Center, and rheumatology experts from the University of Miami, the University of California at San Francisco, and Johns Hopkins University.

Online x-ray archive. The complete set of 17,000 NHANES II x-ray images in the full-resolution form in which they were digitized were made publicly available. These images are available by FTP access to the CEB main network server, and have been accessed by researchers from both within the U.S. and also from international sites. For viewing the x-rays, we created the ImViewJ software, a Java application that may be downloaded from our Web site and which allows the viewing of the images at their full spatial resolutions (1463x1755 for the cervical spine images, 2048x2487 for the lumbar spine images). In addition to the full resolution x-rays, we have also made available a small subset of the images in 8-bit TIFF format, for compatibility with common viewing and processing software. For these images we have included the coordinate data collected under the supervision of a radiologist at Georgetown University. These coordinate data define landmark points for each vertebra commonly used in the field of vertebral morphometry, and serve as reference data to aid in creating and evaluating the performance of image processing algorithms for assisting in the segmentation of the vertebrae. During the current year, we added to this archive TIFF versions of all of the 550 images with radiologist-collected vertebral boundary marks. Users may access this data either through the FTP archive or through the WebMIRS system.

We have given access to this x-ray archive to approximately 80 institutions, of which the last five are: (i) University of Tennessee, Memphis, TN; (ii) University of Maryland School of Medicine, Baltimore, MD; (iii) University of Wisconsin, Madison, WI; (iv) Image Metrics, Manchester, England; (v) NUI Galway, Galway, Ireland.

Image data from this archive was recently used in an image processing technical paper (R. Bernard et al) published in the Proceedings of the Fourth International on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001).

Computer-assisted methods for the extraction of biomedical information from spine x-ray images. The overall goal of this research is to develop methods for better extraction of biomedical information from digital images of the spine. This work has implications both for indexing of image data and for search of that data. For example, for the 17,000 NHANES II images, the only indexing data available is the collateral (alphanumeric) data collected in the questionnaires and examinations; no indexing information derived directly from the images is available, and the high cost of employing radiological experts

to compile such data by physical viewing and interpreting each image makes it unlikely that such information will ever be acquired by purely manual means. These circumstances could be reversed if reliable, biomedically-validated software could produce image interpretations automatically. Even in the more likely case that only semi-automated methods should prove feasible, the reduction in labor costs could be sufficient to allow the creation of databases of significant biomedical information in cases where this is not currently economically feasible. This is the implication of research into computer-assisted image indexing. Computer-assisted image searching is a potential enabler of enhanced information extraction from a database that has already been indexed. The most popular form of this type of search is query by example or a variant, query by sketch. In query by example, the user inputs an image, perhaps by selecting from a set of choices provided by the system, or by providing a completely new image, and queries the database by asking, in effect, "Find records with images like this one", usually with respect to one or more characteristics of the example image, such as shape, histogram, or texture. In query by sketch, the input image is replaced by a sketch by the user, using drawing tools provided by the system. In either case, the system analyzes the input into component features, then searches the images in the database for those with similar features. Results are usually returned as a similarity ranking.

Our work toward building such advanced databases is at the stage of the basic image processing research required to analyze the image contents. We are approaching this problem as a three component process: (1) segmentation; (2) identification; and (3) classification. Segmentation is the step of breaking the image into the constituent anatomical objects of interest, with sufficient reliability and accuracy to take meaningful geometric measurements suitable for the next steps of identification and classification; segmentation involves edge detection and edge linking to form connected boundaries, and may involve modeling of the objects to be segmented, either as simple mathematical shapes, or complex shapes computed from manually segmenting the objects from a sample of images. Identification is the process of labeling a segmented object with an anatomical tag, such as "vertebra C3". Classification is the process of categorizing the objects that we have identified on the normal/abnormal scale for a particular biomedical condition, such as "abnormal for anterior osteophytes."

Using an implementation of the Active Shape Modeling (ASM) algorithm developed at the University of Manchester in the form of a MATLAB Toolkit, we have carried out a preliminary evaluation of the applicability of ASM to the segmentation problem for the cervical spine images. ASM requires the construction of an initial model of the

object to be segmented in an image; we did this by building an "average C3 vertebra shape" by manually collecting C3 boundary data from 20 or more x-rays and using the mean shape resulting from this data. The covariance of the collected data is also used in the ASM algorithm to constrain the output shape to which the iterative ASM process converges. Results of the initial testing were judged to be very promising, and resulted in initiation of collaborative image processing work to achieve a more comprehensive evaluation of the usefulness of ASM and related deformable template methods for the segmentation problem, as well as evaluation of methods for anatomy identification and classification in the spine images. Specific work that has been completed or is under way among collaborators includes: research into vertebrae classification methods at the University of Missouri; research into general deformable templates for the vertebrae segmentation, including use of dynamic programming for optimization at Yale University; research into use of fuzzy logic for edge detection of the vertebrae at Catholic University; and research into application of ASM for segmentation of the vertebrae at Texas Tech University. Each of these collaborators uses the CEB FTP site for downloading test images and radiologist-collected coordinate data.

During the current year, extensive research was conducted toward methods for high level analysis of the x-rays to identify basic anatomical structures prior to detailed segmentation. Several MATLAB algorithms developed for this purpose are described in two technical conference publications for the SPIE [Long and Thoma, 2001], one for the IEEE Symposium on Computer-Based Medical Systems [Long and Thoma, 2001], and an invited article in the *Journal of Electronic Imaging* [Long and Thoma, 2001].

In addition, research into the use of Active Shape Modeling for segmentation of the spine vertebrae was conducted in collaboration with Texas Tech University, and resulted in (a) a technical report describing this method for segmenting the cervical spine vertebrae [Sari-Sarraf et al, 2001], (b) new software for vertebral data point collection, and (c) a technical conference publication for the SPIE [Zamora et al, 2001]. A second phase of this work has begun that includes both the cervical and lumbar spines.

Also during the current year, research was conducted into methods for classifying segmented spine vertebrae for presence/absence of anterior osteophytes, in collaboration with the University of Missouri-Rolla. This work resulted in (a) a technical report detailing the methods used and corresponding results [Stanley, 2001], (b) software used for the testing, and (c) one technical conference publication for the 2001 Rocky Mountain Bioengineering Symposium [Stanley and Long,

2001]. Classification methods evaluated were back propagation neural network, K-means, quadratic discriminant, and Learning Vector Quantization 3.

Finally, in 2001, a prototype system for the retrieval of images based on simple vertebral shape models was developed in collaboration with an engineering student sponsored by the NIH Biomedical Engineering Summer Internship Program (BESIP). This MATLAB program allows the user to specify up to 9 control points and the geometric configuration of these points to define an approximate vertebral shape to search for. The prototype database contains 100 cervical and lumbar images, and will rotate and scale each vertebra in each image to identify the best match to the input shape. Alternatively, the user may specify an example vertebra, and the program will search for the best shape match to the example. Current work is directed toward enhancing this program for more efficient searching and more complex shape specification.

Visible Human Image Database Project

AnatLine

This object-oriented database of Visible Human images, indexed for the male thorax region, was completed, in addition to the tools needed to use the system: VHParser and VHDisplay. The first is for unpacking the data files into its individual components (cross-section images, byte masks, coordinate and label tables, etc.). VHDisplay is for displaying both cross-sectional and rendered images. Also, VHDisplay was augmented with speech synthesis to voice names of anatomic structures, as the images are displayed.

The role of Java programming in this project was highlighted in an invited paper by James Seamans at the Worldwide Java Developer Conference in June 2001 in San Francisco.

AnatQuest

With the goal of providing widespread access to the VH images, to users with low speed connections as well, we are developing a new web interface, AnatQuest. This system allows the user to quickly download selected parts of high-resolution images, and then zoom and navigate over these. All the cross-sections as well as 195 rendered images (some of these from outside sources) may be accessed.

Design considerations: images are converted to tiled TIFF; selective downloading and display of these tiles is implemented by a servlet

engine based on the Java Advanced Imaging API and the Java2D API; anatomical labels are displayed by cursor activation on regions defined by byte-masks and label tables. Research is proceeding toward improving performance, e.g., by trading off displayed tile size vs. lossy image compression.

Next Generation Internet: Infrastructure development and applications

Multilateral Initiative on Malaria in Africa. In an effort to increase bandwidth from the current shared 256 kbps satellite channel among the malaria research sites in Africa, engineering staff participated in reviewing a technical proposal from Intelsat. In addition, to demonstrate video quality that may be expected unless bandwidth is increased, tests were conducted with an ISDN gateway in London on simulating a 2-hop satellite link from Africa. Bandwidths used were 128 and 256 kbps, and video quality was found to be marginal.

A review was conducted on teleconferencing services for the MIM project, and staff contributed to an NSF proposal by NCSA to create a center in Nairobi that is similar to others in Kenya. Specifications were generated for VCON Cruiser 384, a PC-based teleconferencing unit (384 Kbps, H.320, H.323 quality and operation).

Infrastructure. Following staff participation in the planning for the connection between NLM and USUHS main campus and their Simulation Center in Forest Glen, MD, basic connectivity between the main campus and NLM was established, and single mode optical fiber was installed between all three sites.

Experimental work. Test systems installed and used inhouse included: Multi-Router Traffic Grapher which monitors the traffic load on network links, and generates HTML pages containing GIF live visual representations of this data; Iperf network performance measuring tool. For a cross-country test for the Visible Embryo project, NetIQ's Qcheck software was employed to measure memory-to-memory tests between the Armed Forces Institute of Pathology through NLM to the San Diego Supercomputing Center.

NGI meetings. Engineering staff represented NLM at the following meetings: Joint Engineering Team (JET) meetings at the National Science Foundation; Multisector Crisis Management Consortium at NCSA; Network Maryland project at Johns Hopkins University Cancer Center; Internet2 Health Sciences Working Group; NASA-sponsored NGI-Mobile workshop; an NSF sponsored workshop on disaster preparedness that sought to promote collaboration among U.S. and

South American researchers.

Maryland Governor's Task Force on High Speed Networks

In 2001, the Lister Hill Center continued to serve as a federal representative to the Maryland Governor's Task Force on High Speed Networks and the Engineering Advisory Group. The Task Force developed a comprehensive plan for bringing the state's network infrastructure in line with the needs of the 21st century. This plan, completed and presented to the legislature, contains recommendations to: a. combine existing state resources to maximize the state's return on investment; b. use existing state owned fiber where available; c. use current right-of-ways the state possesses to add additional fiber in underserved regions such as the Eastern Shore, Western and Southern Maryland; d. provide equity of access to all regions of the state, and support multiple segments of our society; e. promote collaboration among businesses, educational institutions, governmental bodies and research institutions; f. conduct a select number of high priority pilot projects in health care, business infrastructure development, and state government functions. A major contribution by the Lister Hill Center was made in the development of pilot projects in health care involving remote oncology treatment planning and remote intensive care support.

Turning The Pages

The launch of the TTP system was successfully accomplished March 16 with a transatlantic video conferencing link between NLM and the British Library. In the months preceding this, technical work was done by several Lister Hill Center staff in specifying the system (touchscreen monitors and the Mac computers), identifying the sources and creating the documentation to acquire them, modifying touchscreen monitors to interface with Mac computers, installing and setting up the MacOS preferences and variables for best image quality, installing and testing the equipment in the cabinetry, and testing the video conferencing equipment operating via ISDN lines.

Lister Hill Center hosted IEEE conference in July 2001

The *14th Annual IEEE Symposium on Computer-Based Medical Systems* (CBMS 2001) was held on July 26-27, 2001 at the Natcher Conference Center. 90 peer-reviewed papers were presented, five of them by Lister Hill Center research staff. In addition, special sessions on Receiver Operator Characteristics analysis and NIH Grants Funding were included. This symposium, for which Dr. Thoma and Mr. Long served as General Co-Chairs, was planned in cooperation with Dr. Sunanda Mitra, Texas Tech University, Dr. Ian Greenshields,

University of Connecticut, and Dr. Marina Krol, Mt. Sinai Hospital, among others. Ms. Sheila Levy served as the Local Arrangements chair. The symposium was attended by 130 people.

Engineering Laboratories

The R&D conducted by the Communications Engineering Branch rely on laboratories designed, equipped and maintained by the Branch, as well as content resources that support research.

Document Imaging Laboratory. This laboratory supports DocView, MARS and other research and design projects involving document imaging. Housed in this laboratory are advanced systems to electro-optically capture the digital images of documents, and subsystems to perform image enhancement, segmentation, compression, OCR and storage on high density magnetic and optical disk media. The laboratory also includes high-end Pentium-class workstations running under Windows 2000, all connected by 100 Mb/s Ethernet, for performing document image processing. Both inhouse developed and commercial systems are integrated and configured to serve as laboratory testbeds to support research into automated document delivery, document archiving, and techniques for image enhancement, manipulation, portrait vs. landscape mode detection, skew detection, segmentation, compression for high density storage and high speed transmission, omnifont text recognition, and related areas.

Document Image Analysis Test Facility. Designed, developed and maintained by the Communications Engineering Branch, this off-campus facility houses high-end Pentium workstations and servers that constitute MARS-1 and MARS-2 production systems. While routinely used to produce bibliographic citations for MEDLINE, this facility also serves as a laboratory for research into techniques for autozoning, autolabeling, autoreformatting, intelligent spellcheck and other key elements of MARS. Besides real time performance data, also collected and archived are large numbers of bitmapped document images, zoned images, labeled zones, and corresponding OCR output data. This collection serves as ground truth data for research in document image analysis and understanding.

Image Processing Laboratory. The Communications Engineering Branch Image Processing Lab is equipped with a variety of high end servers, workstations and storage devices connected by 100 Mb/s Ethernet. The laboratory supports the investigation of image processing techniques for both grayscale and color biomedical imagery at high resolution. In addition to computer and communications resources and image processing equipment to capture, process, transmit and display

such high-resolution digital images, the laboratory also has a variety of image content.

The equipment includes a Sun Ultra Enterprise 4000 server with dual 168 MHz CPUs, 512 MB memory, 18 GB of disk storage for application development, and two locally attached Sun StorEdge A5000 RAIDs. Additional computers in the lab include two Sun Ultra 10 workstations, each with a 440 MHz CPU, 512 MB memory and an external 18 GB SCSI disk, two Sun Ultra 10s each with a 330 MHz CPU and 512 MB memory, a Sun SPARCstation 10 with two 40 MHz CPUs and 256 MB memory, and a Sun SPARCstation 20 with two 50 MHz CPUs and 256 MB memory. All of these machines run the Sun Solaris 2.7 operating system.

A Sun SPARCserver 670 machine hosts a 144?platter rewritable magneto?optical jukebox in the lab and runs Sun Solaris 2.5.1. Standalone magneto?optical drives are available in the lab for reading optical platters with capacities of 1X (600 MB) and 4X (2.4 GB). Large?scale magnetic storage is provided by two RAID systems, the larger of which consists of a pair of Sun StorEdge A5000 storage arrays which are attached to the host via dual?looped fibrechannel connections with a maximum throughput of 200 MB/sec, and which provide approximately 150 GB of storage. A small Sun Sparc Storage Array Model 100 RAID system provides an additional 25 GB of storage.

Two ultra-high-resolution E?systems Megascan displays provide image display at spatial resolution of 2048x2560 pixels. An IBM-compatible PC and monitor are also available in the lab for PC testing of Java applications.

Most machines are equipped with multiple networking ports (FDDI, ATM, Ethernet, fast Ethernet) which allow, in addition to standard networking capabilities on the local Ethernet, the capability of alternate physical communications channels with these machines. This capability has been used in communications engineering experiments for point-to-point satellite channels connecting these machines with remote sites. ATM switches connect the Ethernet and FDDI networks to other local area networks throughout the building, to the Internet, and to experimental ATM networks such as ATDnet and MCI's research network, in addition to vBNS, the infrastructure for the Next Generation Internet and Internet-2 initiatives.

Image Processing Lab content resources. A large part of the NHANES II data has been put into the WebMIRS database tables. All of the NHANES II demographic, anthropometric, physical examination, and

adult health questionnaire data is available through WebMIRS, as well as the statistical weighting and sampling strata variables required for analysis of the data. This data covers a sample of approximately 20,000 survey participants. In addition, the 17,000 NHANES II cervical and lumbar spine x-ray images are available for viewing through WebMIRS, in one-quarter spatial resolution format. These 17,000 images are stored on magneto-optical platters accessible through a four-drive 144-platter jukebox hosted by a Sun computer. In addition, all 17,000 images are stored in a magnetic RAID system and are available for public downloading via FTP.

Similarly, a large part of the NHANES III data has been put into the WebMIRS database tables. All of the NHANES III demographic, physical examination, health questionnaire, and laboratory data is available through WebMIRS, as well as the statistical weighting and sampling strata variables required for analysis of this data. The NHANES III data covers a sample of approximately 30,000 survey participants.

In addition to the data above, The Image Processing Lab also contains a selection of History of Medicine color images digitized at high resolution from the Library's Arabic and Persian medical manuscript collection.

Proteus Project

With the goal of developing a system for medical decision making, data entry and data storage in a clinical setting, a Medical Informatics Fellow undertook an investigation of system architecture for using medical knowledge in the form of executable distributed components to construct clinical protocols and thereby to represent the clinical process. In this approach, called Proteus (PROTocols Editable by USers), clinical processes are represented by three types of "knowledge components": actions, processes and events. Each such "KC" has a mechanism to infer its own value and to determine the next action to be launched.

A Java-based proof-of-concept module (Protean) based on the Proteus architecture was built, and as a way to demonstrate its operation, a clinical protocol *Magnesium Sulfate therapy for severe pre-Eclampsia/Eclampsia* was created and demonstrated. This year the editing feature in Protean was improved by (a) creating new activity links between different KCs by just dragging from one KC's connection points to another, and (b) dragging a new KC from a panel graphically displaying different KCs onto the protocol that is loaded

and visible on the main Protean screen.

To identify a suitable inferencing tool, a comparison was made of CTX, the criteria based inferencing tool developed at NLM, and Jess, a rule-based expert system written in Java. Jess is a Java clone of the famous CLIPS of NASA, and it is easier to access, since Protean is also written in Java. Jess also offers a rich set of functionality, which may be exploited at a later stage. These reasons determined the choice of Jess as the inference tool for Proteus.

Also, to avoid the need for a user to create complex rules for every KC created, a "user as an inference tool" feature was incorporated into Protean. When a KC with the user as its designated inference tool, has to make an inference, a dialog box is displayed showing the preconditions on which the decision has to be based along with all the decision-options that are valid in that situation in a combo-box. The user may then select the most appropriate one. As an option, the user could be provided with other support information like web pages or multimedia information that may support the decision-making.

Another role for the user-inference tool is to serve as a fall back mechanism, i.e., if the automated inference mechanisms fail to make a valid inference for any reason, the inference making is passed on to the user. With this capability we have further verified our hypothesis of pluggable inference tools. We can easily switch from one type of inference tool to another by simply selecting the desired one from a combo list.

One benefit of using clinically meaningful entities - the clinical knowledge components - is that new uses, which depend on clinical semantics, can be incorporated with relatively little effort. To demonstrate this aspect of the Proteus approach, some *just-in-time* features were introduced into Protean. The Just In Time project at the Lister Hill Center aims to create generic questions that can be instantiated for specific clinical situations. A key requirement for such an approach is to discover what the clinician is engaged with at any moment while managing a patient. The Proteus approach lends itself to addressing this requirement. If the user selects any transaction KC, a window opens and shows in a tree structure all the possible questions pertaining to the situation represented by the KC, organized into different categories. When the user selects the questions the user is interested in, and clicks on the "answer" button, a browser is opened with PubMed responses to a query string representing the question.

To provide JIT functionality each KC was associated with a "category" - which represents the generic nature of the KC in clinical terms (e.g.,

"drug", "test", "clinical finding"). Also, a "term" is associated with the KC (e.g., the KC dealing with suspected breast lump is associated with term "Malignant Neoplasm of Breast"), which represents the core concern of the KC. This is preferably a MeSH term. To help the user select the appropriate MeSH term, the Protean editor can find the appropriate MeSH equivalent of a term typed in the field by querying the UMLS Knowledge Source Server. The user may also type in a broader term and get a more specific term by clicking on the "narrow" button. This opens another window that shows, as a tree, the broad term and all its sibling terms as branches, with more specific terms shown as leaves.

CEB Publications for calendar year 2001

Long LR, Thoma GR. Identification and classification of spine vertebrae by automated methods. Proc. SPIE Medical Imaging 2001: Image Processing, San Diego CA, February 2001.

Stanley RJ, Long LR. A radius of curvature approach to cervical spine vertebra image analysis. Proc. 38th Annual Rocky Mountain Bioengineering Symposium, 37: 385-90, April 2001.

Zamora G, Sari-sarraf H, Mitra S, Long LR. Estimation of orientation and position of cervical vertebrae for segmentation with active shape models. Proc. SPIE Medical Imaging 2001: Image Processing. Vol. 4322, San Diego, CA, February 2001, 378-87.

Long LR, Thoma GR. Feature indexing in a database of digitized x-rays. Proc. SPIE Storage and Retrieval for Media Databases 2001, San Jose CA, January 2001, 393-403.

Lasko TA, Hauser SE. Approximate string matching algorithms for limited-vocabulary OCR output correction. Proc. SPIE, Vol. 4307, Document Recognition and Retrieval VIII, January 2001, 241-9.

Ford G, Hauser SE, Le DX, Thoma GR. Pattern matching techniques for correcting low confidence OCR words in a known context. Proc. SPIE, Vol. 4307, Document Recognition and Retrieval VIII, January 2001, 241-9.

Kim J, Le DX, Thoma GR. Automated Labeling in Document Images. Proc. SPIE, Vol. 4307, Document Recognition and Retrieval VIII, San Jose CA, January 2001, 111-22.

Long LR, Thoma GR. Computer assisted retrieval of biomedical image features from spine x-rays: progress and prospects. Proc. 14th IEEE

Symposium on Computer-Based Medical Systems. Los Alamitos CA: IEEE Computer Society. July 2001, 46-50.

Tran LQ, Moon CW, Le DX, Thoma GR. Web page downloading and classification. Proc. 14th IEEE Symposium on Computer-Based Medical Systems. Los Alamitos CA: IEEE Computer Society. July 2001, 321-6.

Le DX, Tran LQ, Chow J, Kim J, Hauser SE, Moon CW, Thoma GR. Automated medical records citation records creation for Web-based online journals. Proc. 14th IEEE Symposium on Computer-Based Medical Systems, Los Alamitos CA: IEEE Computer Society. July 2001, 315-20.

Pearson G, Moon CW. Bridging two biomedical journal databases with XML: A case study. Proc. 14th IEEE Symposium on Computer-Based Medical Systems. Los Alamitos CA: IEEE Computer Society. July 2001, 309-14.

Shah H. Proteus: A model for clinical protocols created from Knowledge Components. Proc. 14th IEEE Symposium on Computer-Based Medical Systems. Los Alamitos CA: IEEE Computer Society. July 2001, 59-64.

Krainak DM. A method of content-based image retrieval for a spinal x-ray image database. Poster No. 198, NIH 2001 Summer Research Program Poster Day, August 9, 2001.

Schlaifer JD. OCR affiliations: feasibility considerations and numerical scoring for correction from past datasets. Poster No. 360, NIH 2001 Summer Research Program Poster Day, August 9, 2001.

Invited papers

Seamans J. Visible Human image display program using Java technology. JavaOne: Worldwide Java Developers Conference, Paper TS-1688, June 7, 2001, San Francisco CA. http://servlet.java.sun.com/javaone/conf/sessions/2-0-0/0-sf2001.jsp

Long LR, Thoma GR. Landmarking and region localization in spine x-rays. Journal of Electronic Imaging 10(4), October 2001 (forthcoming).

Extramural research related to CEB work:

Bernard R, Likar B, Pernus F. Segmenting articulated structures by hiera statistical modeling of shape, appearance, and topology, Proc. 4th Interna

Conference on Medical Image Computing and Computer-Assisted Interv
MICCAI 2001, Utrect, The Netherlands, 14-17 October, 2001.

Sari-sarraf H, Mitra S, Zamora G, Tezmol A. Customized active shape models for segmentation of cervical and lumbar spine vertebrae. Texas Tech University College of Engineering Technical Report, available at http://www.cvial.ttu.edu/~sarraf.

Stanley RJ. Boundary feature determination. Internal technical report. University of Missouri-Rolla, August 2000.